AFCAPS-FR-2011-0005

March 6, 2008

D. Matthew Trippe
Karen O. Moriarty
Teresa L. Russell
Shonna D. Waters

Human Resources Research Organization
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314

Prepared for
Kenneth L. Schwartz

AFPC/Strategic Research and Assessment Branch (SRAB)

Air Force Personnel Center
Strategic Research and Assessment
HQ AFPC/DSYX
550 C Street West, Ste 45
Randolph AFB TX 78150-4747

Approved for public release; distribution is unlimited.
UNCLASSIFIED

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report was cleared for release by HQ AFPC/DSYX Strategic Research and Assessment Branch and is releasable to the Defense Technical Information Center.

This report is published as received with minor grammatical corrections. The views expressed are those of the authors and not necessarily those of the United States Government, the United States Department of Defense, or the United States Air Force. In the interest of expediting publication of impartial statistical analysis of Air Force tests SRAB does not edit nor revise Contractor assessments appropriate to the private sector which do not apply within military context.

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information Center - http://www.dtic.mil/

Available for public release. Distribution Unlimited. Please contact AFPC/DSYX Strategic Research and Assessment with any questions or concerns with the report.

This paper has been reviewed by the Air Force Center for Applied Personnel Studies (AFCAPS) and is approved for publication. AFCAPS members include: Senior Editor Dr. Thomas Carretta AFMC 711 HPW/RHCI, Associate Editor Dr. Gregory Manley HQ AFPC/DSYX, Dr. Lisa Mills AF/A1, Dr. Paul Ditullio HQ AFPC/DSYX, Kenneth Schwartz HQ AFPC/DSYX, Johnny Weissmuller HQ AFPC/DSYX, Dr. Laura Barron HQ AFPC/DSYX, Dr. Mark Rose HQ AFPC/DSYX, and Brian Chasse HQ AFPC/DSYX.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE ***(DD-MM-YYYY)*** 06-03-2008	**2. REPORT TYPE** Final	**3. DATES COVERED** ***(From - To)*** September 2007 - September 2008

4. TITLE AND SUBTITLE A Review of Test Item Types	**5a. CONTRACT NUMBER** FA3089-07-F-0462
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Matthew D. Trippe, Karen O. Moriarty, Teresa L. Russell, Shonna D. Waters, Kenneth L. Schwartz, Johnny J. Weissmuller	**5d. PROJECT NUMBER** SPR-99-13
	5e. TASK NUMBER 32
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314	**8. PERFORMING ORGANIZATION REPORT NUMBER** GS-10F-0087 J
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) **Air Force Personnel Center** **Strategic Research and Assessment Branch** **Randolph AFB TX 78150**	**10. SPONSOR/MONITOR'S ACRONYM(S)** HQ AFPC/DSYX
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFCAPS-FR-2011-0005

12. DISTRIBUTION / AVAILABILITY STATEMENT
Available for public release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The purpose of the report is to identify item types and formats that could be included on a test of Information and Communication Technology (ICT) aptitude that would be used in conjunction with and aid in the predictive ability of the Armed Services Vocational Aptitude Battery (ASVAB). The report covers the benefits and drawbacks of various question formats like multiple choice, true/false, etc. It was suggested that multiple choice, information style, logic based reasoning, biodata, and non-verbal reasoning would be the most beneficial item forma for the ICT test.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: **Unclassified**			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Kenneth L. Schwartz
a. REPORT **U**	**b. ABSTRACT** **U**	**c. THIS PAGE** **U**	**U**	30	**19b. TELEPHONE NUMBER** ***(include area code)*** 210-565-3139

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

PAGE LEFT INTENTIONALLY BLANK

A REVIEW OF TEST ITEM TYPES

Contents

Figures

A REVIEW OF TEST ITEM TYPES

I. Background

In 2005-2006, the Defense Manpower Data Center (DMDC) conducted a review of the Armed Services Vocational Aptitude Battery (ASVAB). A panel of experts developed recommendations for changes to the battery, including research that may be necessary to evaluate proposed changes and/or to implement those changes; and prioritized the recommendations in terms of costs, benefits, and timing (Drasgow, Embretson, Kyllonen & Schmitt, 2006).

One of the panel's recommendations stated that *research should be conducted to develop and evaluate a test of information and communications technology literacy. The efficacy of coaching and item familiarity, as well as the feasibility of creating multiple forms, should be examined in conjunction with test development*. Toward this end, Russell and Sellman (2007) reviewed research literature on the assessment of information and communication technology literacy (ICTL). The objectives of that review were to develop a working definition of ICTL based on prior research and to identify and review existing tests of ICTL.

Purpose

In 2007, the U.S. Air Force Personnel Center contracted with the Human Resources Research Organization (HumRRO) to develop and validate a measure of ICT aptitude. That project involved reviewing materials to identify knowledge, skills, and abilities needed for ICT (or cyber) jobs, development of test specifications and test items, pilot-testing the items, and assembling two equivalent forms of the ICT test. In addition to the pre-equated test forms, several interim reports were produced on particular topics related to the test development. This report is one of the interim project reports.

The objective of this report was to define item types and formats that could be included on a test of ICT aptitude. This review will aided development of test specifications for the ICT measure. The report describes several test item types and, where appropriate, our experience with that item type and relevant research. For each item type, information is provided about the capabilities of an item development and banking software program that we planed to use—Perception™, owned by Question*mark*. Information about this software including costs, manuals, and software tryout is available on their website, http://www.questionmark.com/us/index.aspx. Specifically, we will discuss the capabilities of Perception™ Version 3 software with respect to various test item types. Version 4 is the current version and has more capabilities than Version 3. However, we were not be able to upgrade to Version 4 in time for the pilot testing which occurred in the summer of 2008. Hence our focus was on Version 3.

II. Multiple Choice Test Items

Multiple choice (MC) questions have a stem which presents the problem and several response options. The examinee chooses an answer from the response options. The benefits of multiple choice items are that they require relatively few resources to develop, can be administered and scored relatively easily and can cover a wide breadth of content. The drawbacks of multiple choice items are that they can have poor face validity, are difficult to write to assess higher level thinking, and that the probability of guessing the correct answer can be non-negligible. Computer based tests allow for extended multiple choice items, which are similar to traditional multiple choice items, but the number of answer choices is large enough that the probability of guessing correctly is very low. For example, examinees may be presented with a paragraph and asked to highlight the sentence or word that addresses the stem (Sireci & Zenisky, 2006).

The MC format is widely used to assess a variety of individual difference constructs, and several special types of MC items have evolved. Some formats that could be useful for the information and communications technology aptitude test include:

- Information
- Logic-based reasoning
- Situational judgment
- Non-verbal reasoning
- Scenario/stimulus-based test
- Biographical data

Information

Information tests are a special class of declarative knowledge tests. If knowledge tests were placed on a continuum ranging from everyday or general knowledge to highly specialized job knowledge, information tests would anchor the low end. They measure knowledge that anyone interested in a particular topic might learn from their choices of recreational and educational pursuits. The key notion is that information tests are surrogate measures of motivation and skill in a particular area.

Information tests were among the most successful and most highly valid printed classification tests created by the Army Air Forces Aviation Psychology Program during World War II. Guilford and Lacey (1947) described the logic of information tests as follows:

> It is becoming recognized more and more that what a person knows or does not know can be used to reveal a number of things concerning his personal background. Since he is to a large extent a product of his personal experience, and since what he is bodes good or ill concerning his future status in one respect or another, knowledge scores promise to have predictive value (p. 341).

AAF researchers defined knowledge likely to transfer to piloting, thought to be indicators of aviation interest, and expected to indicate skills relevant to aircrew jobs. One successful test was the Technical Vocabulary Test. Some items had to do with planes, others with maps or astronomy. Some example items from this test appear in Figure 2.1. Other tests had items testing very fundamental knowledge of sports that AAF researchers expected would build skills related to piloting.

The plane with a cannon in its nose is manufactured by:

A. Bell.
B. Boeing.
C. Sikorsky.
D. Douglas.
E. Vultee.

Time is usually calculated with reference to:

A. The Naval Observatory in Washington.
B. Zero degrees latitude.
C. Greenwich.
D. The International Date Line
E. The League of Nations' Observatory in Geneva.

Figure 2.1. Example AAF Technical Vocabulary Test items (Guilford & Lacey, 1947).

Information tests continue to serve military selection and classification well today. General Science (GS), Electronics Information (EI), and Auto and Shop Information (AS) are all basically information tests.

Information items are good candidates for inclusion on the ICT aptitude test. After decades of use, they have proven successful for use in military selection and classification. They are likely to be very useful predictors of performance in training for cyber jobs. The key would be to define knowledge that youth with high ICT aptitude are likely to have learned by searching the Internet, pursuing spare time activities, and taking courses in high school.

Logic-Based Reasoning

Logic-based reasoning (LBR) items assess inductive or deductive reasoning skills by presenting examinees with a premise or set of premises and asking them to choose the one valid conclusion among a series of conclusions (Colberg, Nester, & Trattner, 1985). Deductive LBR items are essentially formal syllogisms placed in the scaffolding of a traditional verbal reasoning test item. Inductive LBR items are similar in structure, but rely on probabilistic rather than necessary premises and conclusions. An example LBR item appears in Figure 2.2. This affords assessment of verbal reasoning to have objective qualities comparable to mathematics assessments (Colberg, 1985). That is, the correct answer represents a necessary inference and distracters represent necessarily incorrect inferences with the same precision as traditional mathematics items. In contrast, traditional verbal reasoning items often rely on informal inferences that may be ambiguous and can lead to alternative subjective or idiosyncratic interpretations that are plausible (or at least not necessarily incorrect). Traditional number or figure series items are subject to the same criticism (Colberg et al., 1985).

Police officers were led to believe that many weapons sold at a certain gun store were sold illegally. Upon investigating the lead, the officers learned that all of the weapons sold by the store that were made by Precision Arms were sold legally. They also found that none of the illegally sold weapons were .45 caliber.

From the information given above, it can be validly concluded that, concerning the weapons sold at the store,

A) all of the .45 caliber weapons were made by Precision Arms
B) none of the .45 caliber weapons were made by Precision Arms
C) some of the weapons made by Precision Arms were .45 caliber weapons
D) all of the .45 caliber weapons that were sold were sold legally
E) some of the weapons made by Precision Arms were sold illegally

Figure 2.2. Example logic-based reasoning item.

Another desirable aspect of LBR items is that form equivalence can be virtually guaranteed. Because syllogisms or proofs are the basis of LBR items, the same fundamental set of premises can be used repeatedly in differing item contexts. In other words, the same premise (e.g., all S are P) can be fitted to almost any item stem (e.g., all disks are copy protected) to form items that have different facades but measure the same reasoning construct. Moroever, there are existing structural taxonomies of LBR items, complete with valid and invalid conclusions (Colberg, 1984; Colberg & Varon Cobos, 2000; Simpson & Nester, 2007). LBR items are resource intensive to develop, but a higher than normal proportion of items survive the pilot procedure, perhaps due to the systematic structuring of the items.

Colberg et al. (1985) argued that from a psychometric perspective, deductive and inductive LBR measures are convergent and need not be treated as separate measures. Analysis of two separate samples revealed correlations corrected for unreliability of .90 and .99. It should be noted that the LBR measures in these analyses were relatively short and unreliable and thus the correction was substantial (uncorrected correlations were .46 and .43). Another noteworthy finding was that both LBR measures were also highly correlated with measures of reading comprehension (r = .87 corrected for unreliability for both inductive and deductive measures) and arithmetic reasoning (r = .65 for deductive and .71 for inductive, corrected for unreliability). This finding has relevance for LBR measure’s potential to provide incremental validity over the existing Armed Service Vocational Aptitude Battery (ASVAB). Even so, the LBR format is a useful one for getting at algorithmic thinking and seems to parallel the kind of aptitude needed in many ICT jobs very well. It could be useful to find a way to minimize the verbal load of the items, perhaps through the use of graphics or symbols. This might also reduce some of the overlap with the ASVAB.

Situational Judgment

Situational judgment tests (SJTs) have become increasingly popular in employment testing in recent years because they (a) address knowledge and skills that are difficult to measure with traditional multiple-choice test formats, (b) yield reasonably high estimated validities for predicting job performance (average *r* = .34 uncorrected) and incremental validity over general cognitive ability ($\Delta r =$ .08 corrected); (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001) and (d) typically yield small to moderate subgroup differences (Hough, Oswald, & Ployhart, 2001).

SJTs provide a verbal or written description of a scenario and a list of potential actions that could be taken. An example appears in Figure 2.3. In some instances, the respondent reads the situation and indicates (a) which action he/she believes is *most* effective and (b) which action he/she believes is *least* effective (Weekley & Jones, 1999). Other formats have asked the respondent to indicate what he or she would be most and least likely to do in the situation (Motowidlo, Dunnette, & Carter, 1990) or to rate the effectiveness of several actions (e.g., Waugh & Russell, 2005).

You are a flight attendant on a plane. You have just started telling the passengers the safety procedures. One of the passengers says, in a loud voice, to his traveling companion that people who listen to the safety instructions are wasting their time because plane crashes are so rare. He then continues to talk loudly to his friend and ignores you. What would you do?

a. Explain to the passenger that although plane crashes are rare, it is important to be prepared.
b. Ask the passenger to be quiet or he/she will be removed from the plane.
c. Talk over the passenger in a louder voice.
d. Whistle loudly to get everyone's attention. Then tell everyone to be quiet while you are giving the safety instructions.

Figure 2.3. Example situational judgment item.

Tests using an SJT format have been around for more than 100 years (Weekley & Ployhart, 2006). The primary debate then and today has to do with what SJTs measure. One point of agreement is that an SJT is a measurement method—a format of a test. What it measures is a function of content choices made by developers. At the highest level, SJTs simply measure judgment (Schmitt & Chan, 2006). Virtually all SJTs have a strong interpersonal component, and some SJTs have a positive relationship with cognitive ability. Some examples of constructs that SJTs have been constructed to assess include conflict resolution (Drasgow, Olson-Buchanan, & Moberg, 1999), managerial skills (Motowidlo, Hanson, & Crafts, 1997), and even technical skills (Hanson, Borman, Mogilka, Manning, & Hedge, 1999). While technically-oriented SJT items could be constructed for ICT measurement, the items would probably need to be too job-specific for use in entry-level selection and classification. We expect that this format could be useful for higher level cyber jobs that require providing technical advice and information to commanders or working as team.

Non-Verbal Reasoning (NVR)

Carroll (1993) argued that there are three first-order factors related to the domain of Reasoning: RG (Sequential Reasoning), I (Induction), and RQ (Quantitative Reasoning). The Raven’s Progressive Matrices test is an example of an Inductive Reasoning measure. Carroll also noted that Visualization (VZ) is often related to reasoning abilities in that it involves the ability to apprehend, encode, and mentally manipulate spatial forms (p. 309). An example of a test that would also load on a VZ factor is Paper Folding. This is considered distinct from Spatial Relations (SR), which involves simpler speeded tasks involving cards, figures, and flags. Thus, reasoning tests may have a VZ component as long as the requirement to mentally manipulate spatial forms is accompanied by an inductive, deductive, or sequential reasoning task. In addition, reasoning tests should be more reliant on power than speed.

NVR tests can be more generally thought of as a subset of reasoning measures. For the purpose of this review, NVR is the ability to identify patterns, apply rules, and draw inferences in problems presented visually. It is measured by tests that require little or no reading but instead rely on pictures, figures, symbols, and/or geometric patterns. NVR tests vary in terms of whether they include verbal instructions; however, they usually include very simple example item(s) so that the test taker can understand the requirements of the task without verbal instructions.

A number of standard item formats are used in measures of NVR. Among these are: pictorial oddities, faulty pictures, figure analogies, spatial analogies, figure series, figure classification, figure generalization, figure matrices, embedded figures, gestalt completion, reversed figures, block counting, cube comparisons, surface development, spatial visualization, object-aperture test, perspective reasoning, paper folding, and figural reasoning matrices (cf., Jensen, 1980). The most common formats among these are figure series and figure matrices. Example items are presented in Figures 2.4 – 2.6.

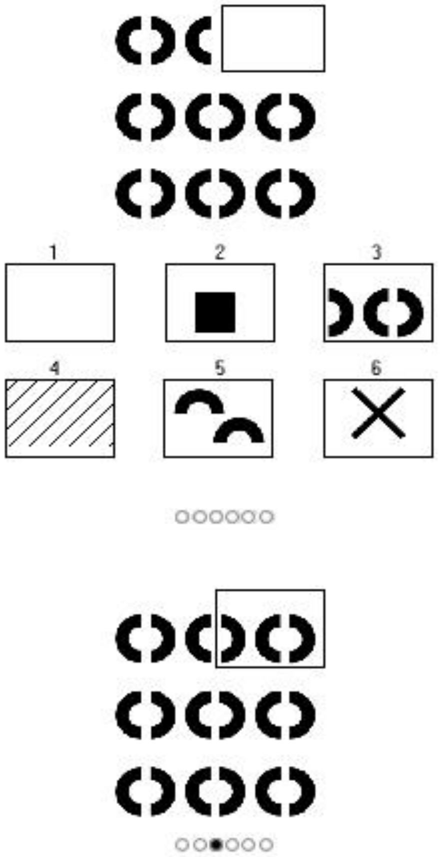


Figure 2.4. Example figure matrices item.

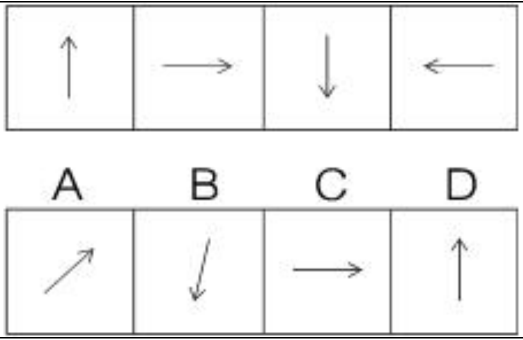


Figure 2.5. Example figure series item.

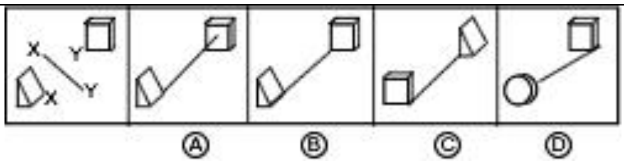
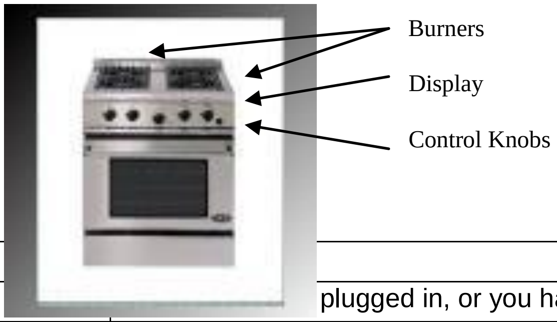


Figure 2.6. Example assembling objects item.

We think NVR is likely to be highly relevant for cyber jobs. DMDC is currently sponsoring a project to assess the construct validity of NVR measures. The project will involve administering one or more marker tests for NVR (i.e., a test with a very strong research track record), along with experimental NVR measures, probably like those in Figures 2.4 and 2.5. Note that the format in Figure 2.6 is that of Assembling Objects (AO), one of the tests currently on the ASVAB. As this project unfolds, we will coordinate with the Defense Manpower Data Center (DMDC) to maximize efforts across the two projects.

Scenario/Stimulus-Based

Scenario- or stimulus-based MC tests present a scenario/stimulus and ask the examinee to respond to several items that require reference to the scenario or stimulus. Reading comprehension tests often have this format, where a reading passage is followed by several multiple choice items about it. Also, tests of graph and table reading often use this format. Figures 2.6 and 2.7 provide examples. In both cases, the stimulus presents troubleshooting rules. The test items require the examinee to apply the rules presented in the stimulus object to new situations.

Stovetop Troubleshooting Guide	
 	
Problem	
Burner fails to light.	plugged in, or you have blown a fuse.
Burner makes popping noise.	The burner is wet from washing. Let it dry.
Burners spark continuously.	There is a faulty spark module. Contact a service technician to replace the module.
The display is showing "PF"	There has been a power failure. Reset the clock.
The control knob will not turn.	You are not pushing in before turning.

1. You have just washed the burner. According to the information in the *Troubleshooting Guide,* what is likely to happen?

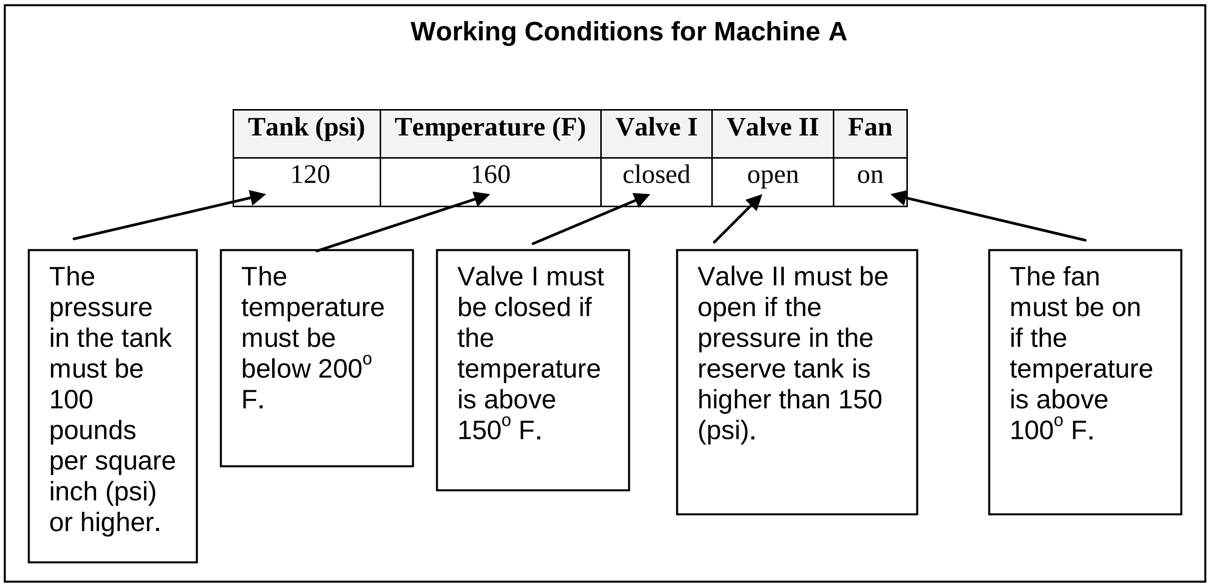
   A. The burner will fail to light.
   B. The burner will make a popping noise.
   C. The burner will spark continuously.
   D. The burner will be unsafe.

2. The burners on the stovetop keep sparking, even when you are not trying to light them. According to the information in the *Troubleshooting Guide*, what should you do?

   A. Check to be sure the stove is plugged in.
   B. Let the burner dry.
   C. Contact a service technician.
   D. Reset the clock

Figure 2.6. Example stimulus-based multiple choice test item.

The main advantage of scenario/stimulus-based items is that they allow for greater interpretation or manipulation of information. There are a couple of drawbacks, however. Development can be more complicated since the stimulus and items are related. Changes to the stimulus can affect all items. The items take a little more testing time than most MC items. For that reason, it is desirable to ask as many questions as possible about one stimulus, particularly in the pilot stage, since some items will not survive pilot testing. If only one or two items survive for a stimulus, it is questionable whether the stimulus set is worth including on the final exam, given time requirements. If something is wrong with one stimulus, all of the items embedded in it will also fail.



Working Conditions for Machine A

Tank (psi)	Temperature (F)	Valve I	Valve II	Fan
120	160	closed	open	on

The pressure in the tank must be 100 pounds per square inch (psi) or higher.

The temperature must be below 200° F.

Valve I must be closed if the temperature is above 150° F.

Valve II must be open if the pressure in the reserve tank is higher than 150 (psi).

The fan must be on if the temperature is above 100° F.

Current Conditions for Machine A				
Tank (psi)	**Temperature (F)**	**Valve I**	**Valve II**	**Fan**
120	180	closed	open	off

1. Which of the following is the source of the problem?

A. High temperature.
B. Low temperature.
C. Fan off.
D. Valve I closed.

Current Conditions for Machine A				
Tank (psi)	**Temperature (F)**	**Valve I**	**Valve II**	**Fan**
80	160	closed	open	off

2. Which of the following is the source of the problem?

A. Low pressure.
B. Low temperature.
C. Fan off.
D. Valve I closed.

Figure 2.7. Another example stimulus-based multiple choice test item.

Biographical Data (Biodata)

Biodata items (Stokes, Mumford, & Owens, 1994) are based on the notion that the best indicator of future performance is past performance (Wernimont & Cambpell, 1968). The idea is that people engage in particular behavioral patterns overtime and that these experiences provide meaningful input to the development of the self (i.e., personality traits); particularly during certain time periods in one's life that are especially conducive to the development of the self concept (e.g., high school years). Salient life experiences (typically negative life experiences) help to shape the current self because of how one has had to adapt to the situation. Biodata items usually assess biographical information relevant to job performance. Past research has indicated that well-constructed biodata measures can exhibit good levels of criterion-related validity (e.g., Bliesener, 1996; Carlson, Scullen, Schmidt, Rothstein, & Erwin, 1999; Dean, 2004; Gandy, Dye, & MacLane, 1994; Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990) and small subgroup differences (e.g., Gandy et al., 1994; Reilly & Chao, 1982). Examples of biodata items appear in Figure 2.8.

In your present job, how often have you typically been late for work?
a. four or more times per week
b. two or three times per week
c. once per week
d. never

Which of the following have you ever experienced or won? (mark all that apply)
a. elected to a leadership position (e.g., class president)
b. team spirit award
c. good attendance award
d. member of an academic honor society
e. citizenship award
f. athletic competition award

Figure 2.8. Example biodata items.

Biodata items have two key characteristics:

1. People are asked to recall and report prior behavior and experiences
2. Questions refer to behavior and experiences occurring in specific situations to which individuals are likely to have been exposed

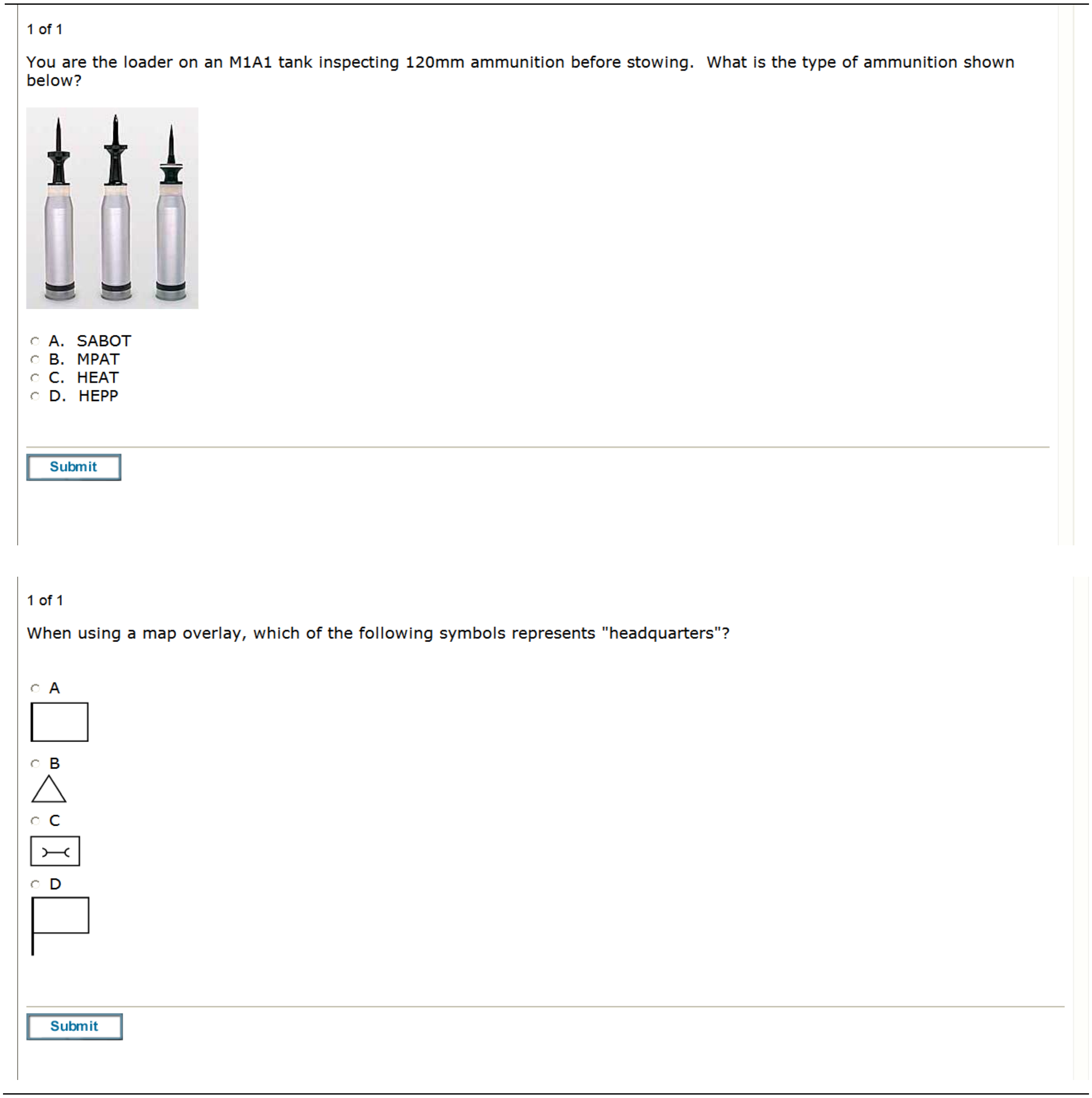
Some ICT-relevant behaviors that could be transformed into biodata items include:

- Participate in virtual environments (e.g., Second Life).
- Play Internet games (e.g., World of Warcraft, Halo).
- Read online publications (e.g., *Wired, PC Gamer).*
- Use instant messaging.
- Learn skills from "how-to" sites on the Internet.
- Use or adapt others' code for own purposes.
- Create their own code for an intended purpose.
- Build their own computers.

- May take high school computer science courses, but those courses may be below their skill level.

Perception Version 3 Capability

Multiple choice items are extremely easy to develop in Perception. As seen in Figure 2.8, graphics can be used as part of the stem or as response options. Scoring is straight forward. When setting scores in Perception, one can select any value wanted. For example, instead of the traditional award of 1 point, the test developer can award 2, 3…*n* points for a correctly selected option.



1 of 1

You are the loader on an M1A1 tank inspecting 120mm ammunition before stowing. What is the type of ammunition shown below?

○ A. SABOT
○ B. MPAT
○ C. HEAT
○ D. HEPP

1 of 1

When using a map overlay, which of the following symbols represents "headquarters"?

○ A

○ B

○ C

○ D

Figure 2.9. Examples of graphics items produced in Perception 3.

III. Non-Traditional Formats

Like multiple choice, non-traditional formats can be used to measure a variety of traits including information, non-verbal reasoning, biographical data and so on. This section describes the following non-traditional formats:

- Multiple response
- Matching
- True-False
- Completion
- Short/extended response
- Drag and drop/Drag and connect
- Point and click
- Performance
- Simulation

Multiple Response

Definition and characteristics

Multiple response items ask examinees to select any number of correct responses to a question. An example appears in Figure 3.1. Multiple response items are subject to many of the same benefits and drawbacks enumerated for multiple choice items. That is, items are relatively easy to develop and score. Nevertheless, face validity and the capability to assess higher level thinking are concerns with such items.

Perception Version 3 capability

Multiple response items are very easy to develop in Perception and scoring is flexible. For instance one can award a non-zero score only when all selections are correct (i.e., test takers select all they should and do not select what they should not). With this option, a test taker earns either all the points or none. Alternatively, one can give a non-zero score for each correctly selected (and non-selected) option up to a maximum of *n* points.

The issue of weighting (or rescoring or rescaling) arises with non-traditional items. One can argue that these items are worth more than traditional, multiple choice items. If so, then how much more? Assume a multiple response item with five response options where partial scoring is allowed. Should this item be worth five times more than a multiple choice item? Does the content covered warrant this, or does that overweight the item? Should we rescale it to be worth only 3 points? If we should rescale, how do we determine the final worth of the item? Empirically? SME judgment? Alternatively we could reject the argument that these items are worth more than multiple choice items, which obviates the need to weight or rescale them.

Matching

Definition and Characteristics:

Matching items ask examinees to pair stimuli in order to address the item stem. The primary benefit of matching items is that they can assess a large amount of content in a relatively compact manner. Additionally, examinees often report enjoying a break from the more common multiple choice format. Scoring matching items is straightforward, but weighting issues described in the multiple response section apply.

Perception Version 3 capability:

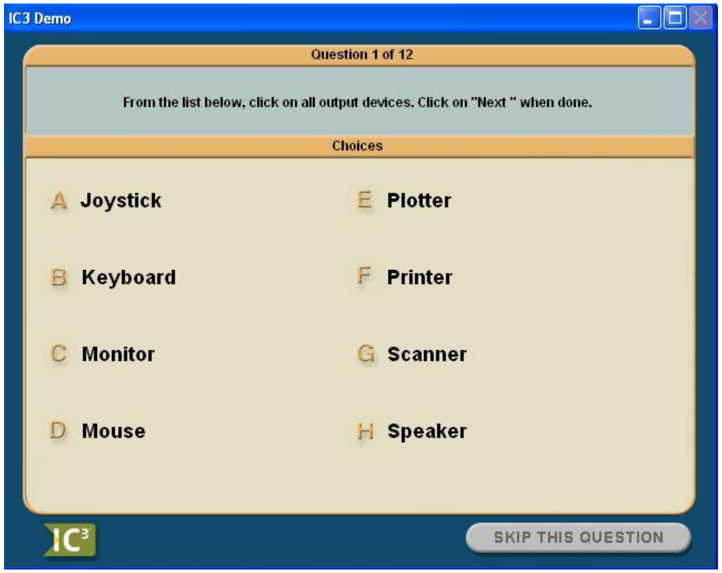
For the most part, these items are easy to create. Matching items can be one of three types. See Figure 3.2 for examples. We can create an item where there are as many cities as there are states and each city can only be selected once. Or, we can create an item where there are more cities than states and each city can only be selected once, as is shown in the left panel of Figure 3.2. Finally, we can create an item like that shown in the right panel of Figure 3.2. Note that these response options must be selected more than once.

Which of these adults has a blood pressure reading outside the normal range?

Select all that apply.

Adult	**BP Reading**
Mary	**150/90**
Bob	120/70
Ellen	100/60
Frank	100/50
Pat	90/50

a. Mary
b. Bob
c. Ellen
d. Frank
e. Pat



IC3 Demo

Question 1 of 12

From the list below, click on all output devices. Click on "Next " when done.

Choices

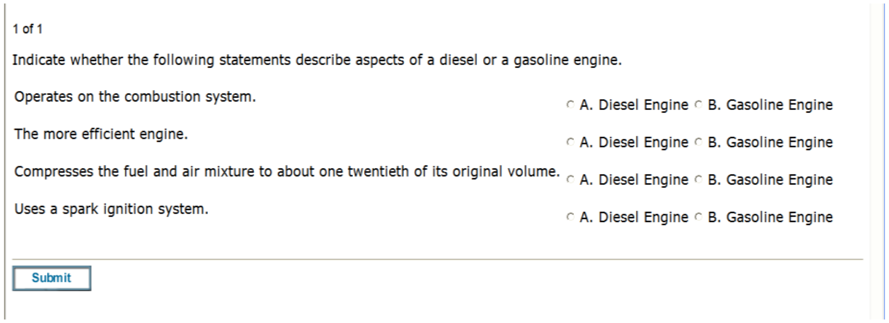
A Joystick
B Keyboard
C Monitor
D Mouse
E Plotter
F Printer
G Scanner
H Speaker

Figure 3.1. Example multiple response items.

Match the following states to their capitals.

1. South Carolina ___
2. Kentucky ___
3. Massachusetts ___
4. Connecticut ___

a. Raleigh
b. Frankfort
c. Fargo
d. Hartford
e. Boston
f. Columbia
g. Pierre



1 of 1

Indicate whether the following statements describe aspects of a diesel or a gasoline engine.

Operates on the combustion system. ○ A. Diesel Engine ○ B. Gasoline Engine

The more efficient engine. ○ A. Diesel Engine ○ B. Gasoline Engine

Compresses the fuel and air mixture to about one twentieth of its original volume. ○ A. Diesel Engine ○ B. Gasoline Engine

Uses a spark ignition system. ○ A. Diesel Engine ○ B. Gasoline Engine

Figure 3.2. Example matching items.

True-False

Definition and Characteristics:

Examinees are asked to evaluate if a given statement is true or false. These items are relatively easy to develop, administer and score. A broad range of content can be covered with true-false items. True-false items tend to encourage guessing and can have poor face validity. As with other formats described so far, it is difficult to assess higher level thinking with this item format.

Perception Version 3 capability:

True/false items are very easy to develop in Perception. Essentially, they are a sub-type of multiple choice items.

> *MS Access is a word processing application.*
> *a True*
> *b. False*

Figure 3.3. Example of a true/false item.

Completion

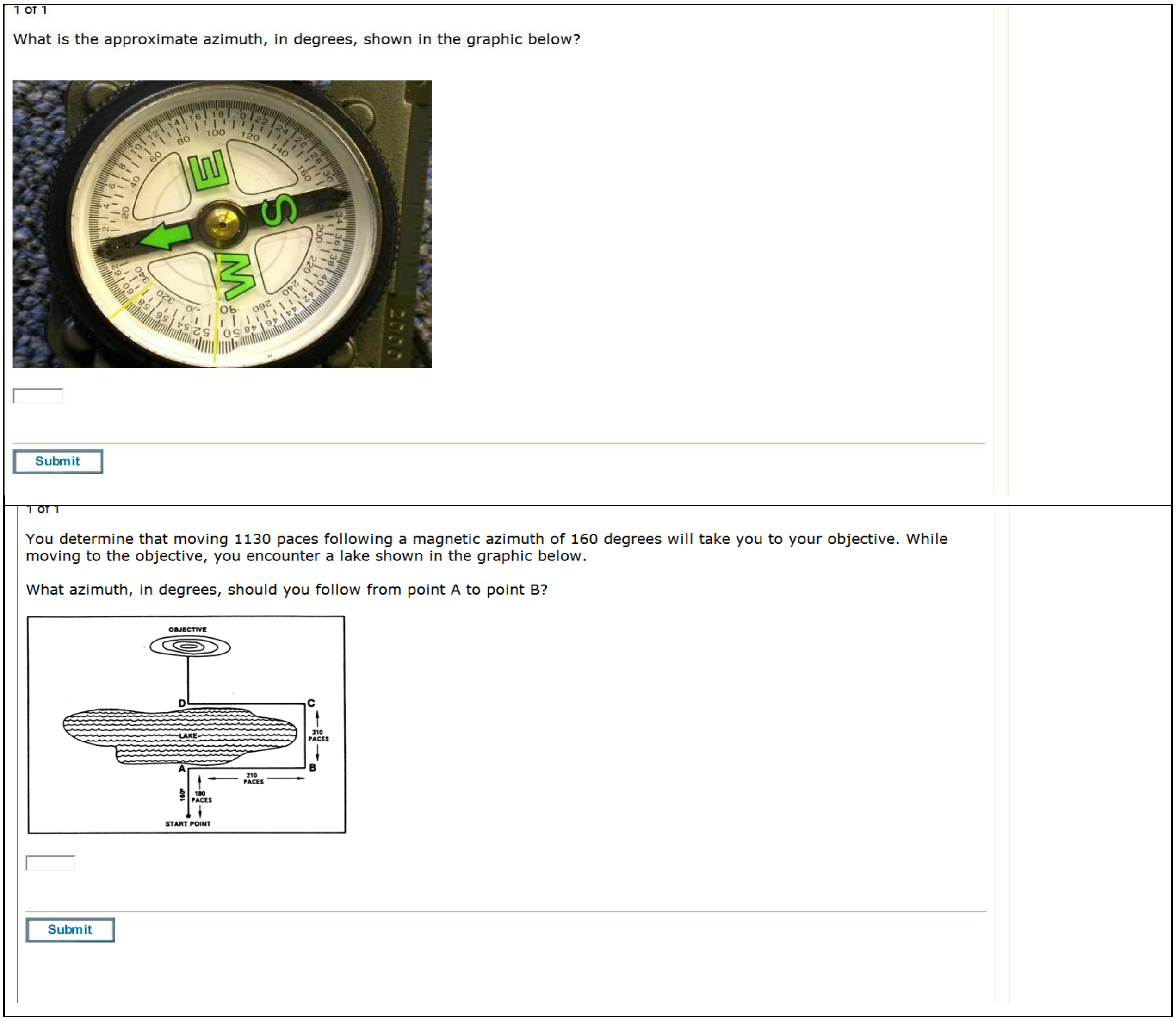
Definition and characteristics:

Simple completion items involve short examinee generated answers. The probability of guessing correctly is much lower in completion items that it is for any item discussed thus far. In addition, completion items can assess recall as opposed to recognition. Although completion items can be scored automatically, scoring is not as straightforward because of the potential for variants of the correct response (synonyms, spelling variants, numerical equivalents). For text completion or fill in the blanks the synonyms issue is very real. Issues with spelling can be very real, too. There are some items where correct spelling is part of the knowledge being measured, but there are also items where spelling is not important. For example, imagine an item where the correct answer is “parallel.” We would have to allow for all reasonable spellings of parallel. Consider if the answer was a phrase, rather than one word. We would have to create a very complicated scoring scheme.

Perception Version 3 capability:

Completion, or fill-in-the-blank, items are easy to create in Perception. Figure 5 shows two examples of numerical fill-in-the-blank items. We can require a specific answer or allow a range of correct responses. These tend to be easier for numerical than text items because we can reduce the number of possible correct answers. Because of the concerns noted above, items requiring fractions should be avoided. Also, if possible, the answers should be multiples of 5 or

10. Text fill-in-the-blank items require very flexible scoring schemes to allow for common misspellings. Perception automatically will check for, and allow (if necessary), credit for common misspellings, but there usually is a need for additional programming.



1 of 1

What is the approximate azimuth, in degrees, shown in the graphic below?

1 of 1

You determine that moving 1130 paces following a magnetic azimuth of 160 degrees will take you to your objective. While moving to the objective, you encounter a lake shown in the graphic below.

What azimuth, in degrees, should you follow from point A to point B?

__Figure 3.4. Examples of completion items.__

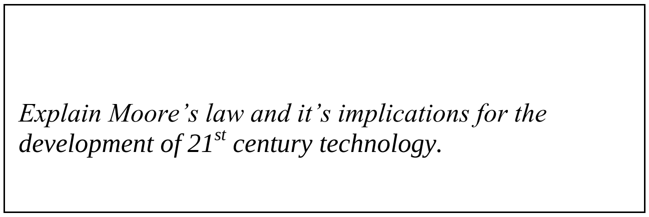
Short/Extended Response

Definition and characteristics:

Examinee are asked to generate a written responses of either short (1-2 sentence) or extended (a paragraph or more) length. These items have greater capacity to assess higher level thinking. Nevertheless, automated scoring of text responses is a new and controversial technology in large scale assessment. Items of this sort are traditionally hand scored, which is resource intensive. Moreover, these items take relatively more time to administer and are memorable to examinees, which compromises test security.

Perception Version 3 capability:

Perception allows Essay items, but they must be scored by hand.



Explain Moore's law and it's implications for the development of 21st century technology.

Figure 3.5. Example of short/extended response items.

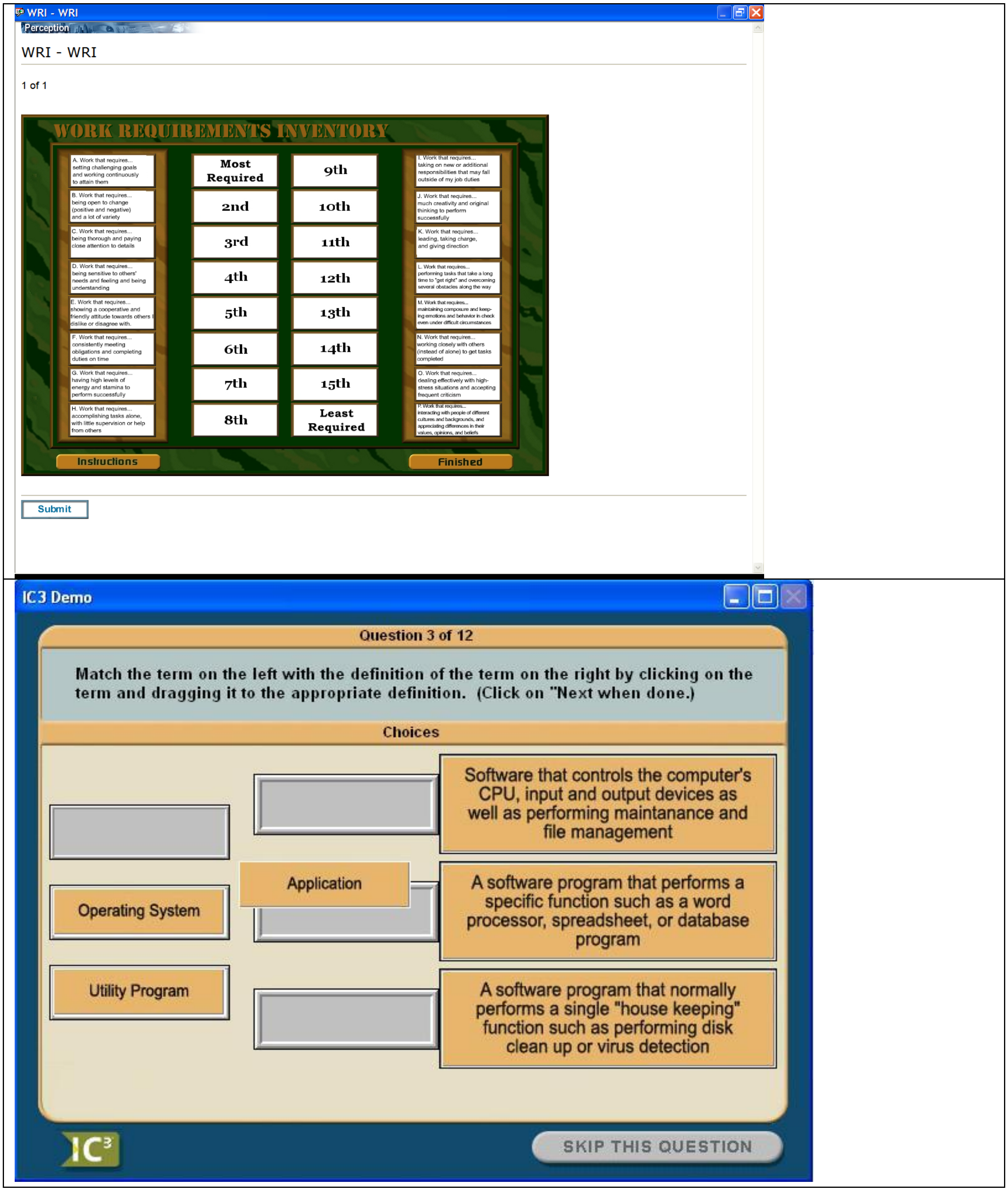
Drag and Drop/Drag and Connect

Definition and characteristics:

Drag and drop items can be used for several tasks. Many are simply enhanced versions of item types already discussed. For example, examinees can accomplish a matching task by dragging a word and dropping it next to its definition. Examinees can also be asked to sort, order, or classify information or to specify relationships among items. The primary advantage of this item format is that it can cover a relatively large amount of content in a single item. Weighting of item scores and face validity are of concern.

Perception Version 3 capability:

Perception has this capability. However, there have been problems with these item types since our operating system was updated to Windows XP™. We suspect there is a conflict, but Question*mark* does not support Version 3 anymore so we are unable to get technical support.



A. Work that requires... setting challenging goals and working continuously to attain them

B. Work that requires... being open to change (positive and negative) and a lot of variety

C. Work that requires... being thorough and paying close attention to details

D. Work that requires... being sensitive to others' needs and feeling and being understanding

E. Work that requires... showing a cooperative and friendly attitude towards others I dislike or disagree with.

F. Work that requires... consistently meeting obligations and completing duties on time

G. Work that requires... having high levels of energy and stamina to perform successfully

H. Work that requires... accomplishing tasks alone, with little supervision or help from others

I. Work that requires... taking on new or additional responsibilities that may fall outside of my job duties

J. Work that requires... much creativity and original thinking to perform successfully

K. Work that requires... leading, taking charge, and giving direction

L. Work that requires... performing tasks that take a long time to "get right" and overcoming several obstacles along the way

M. Work that requires... maintaining composure and keeping emotions and behavior in check even under difficult circumstances

N. Work that requires... working closely with others (instead of alone) to get tasks completed

O. Work that requires... dealing effectively with high-stress situations and accepting frequent criticism

P. Work that requires... interacting with people of different cultures and backgrounds, and appreciating differences in their values, opinions, and beliefs

Match the term on the left with the definition of the term on the right by clicking on the term and dragging it to the appropriate definition. (Click on "Next when done.)

Software that controls the computer's CPU, input and output devices as well as performing maintanance and file management

A software program that performs a specific function such as a word processor, spreadsheet, or database program

A software program that normally performs a single "house keeping" function such as performing disk clean up or virus detection

Figure 3.6. Example of drag and drop items.

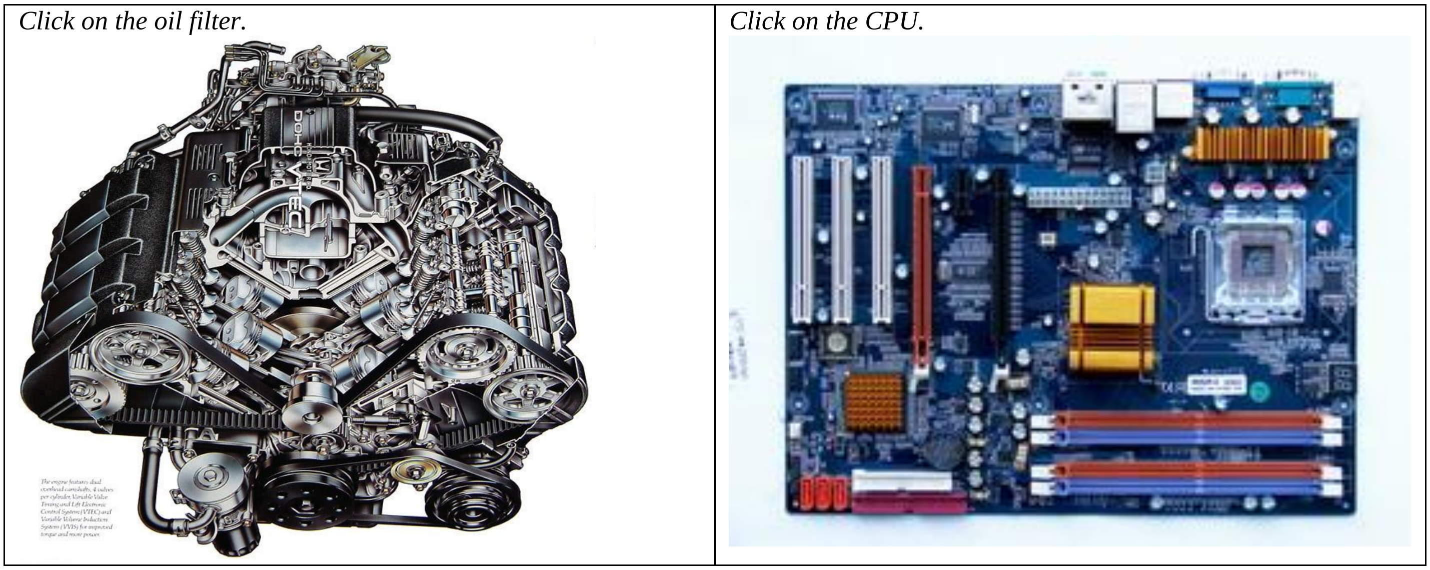
Point and Click

Definition and characteristics:

This is another enhanced matching format. In this format, the examinee is presented with a diagram of some kind. The examinee's task is to point to various portions of the diagram to identify the "stem objects." This item format has face validity and can assess a great deal of knowledge in a single item. Nevertheless, the content that can be assessed in this format is limited.

Perception Version 3 capability:

With some minor editing, this item type could be created in Perception Version 3. Instead of asking the examinee to click on the oil filter or CPU, we could require them to drag and drop an X on the oil filter or CPU. However, as noted above, we are unable create drag and drop items in Perception Version 3.



Click on the oil filter.

Click on the CPU.

Figure 3.7. Example of point and click items.

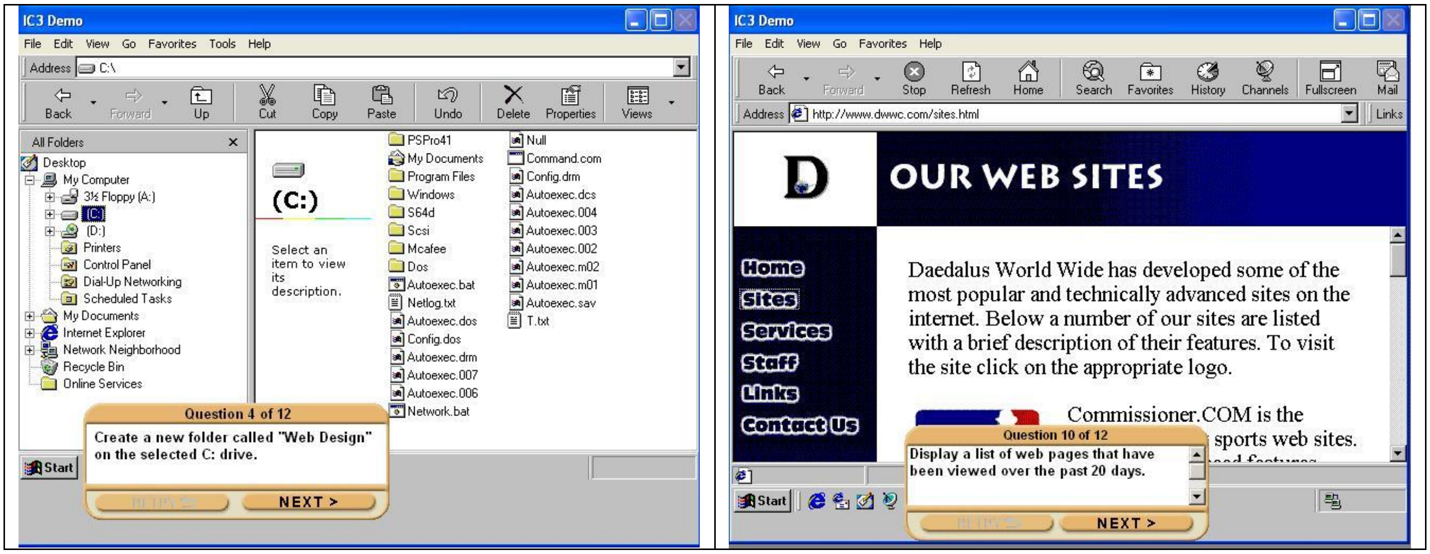
Performance Task

Definition and characteristics:

Performance task items generally test procedural knowledge within an application. That is, the examinee is asked to perform a task virtually as they would in an applied context. The benefit of these items is that they directly assess procedural knowledge and have high face validity. The primary drawback is that programming is required for development.

Perception Version 3 capability:

Version 3 does not have this capability.



Question 4 of 12

Create a new folder called "Web Design" on the selected C: drive.

Daedalus World Wide has developed some of the most popular and technically advanced sites on the internet. Below a number of our sites are listed with a brief description of their features. To visit the site click on the appropriate logo.

Commissioner.COM is the ... sports web sites.

Question 10 of 12

Display a list of web pages that have been viewed over the past 20 days.

Figure 3.8. Example performance task items.

Simulation

Definition and characteristics:

Simulation items are performance task items that are extensive and complex. These items often involve a multi-step response that may include the integration of information across multiple applications and data sources. These items have high face validity and can measure higher level thinking. Development of these items is resource intensive. Scoring is similarly complicated.

Perception Version 3 capability:

Version 3 does not have this capability.

IV. Discussion

Several item formats are potentially useful for the current project. For multiple choice items, information, LBR, biodata, and NVR item types might be particularly useful. Some of the non-traditional formats cannot be accomplished with the software we plan to use. Even so, a number of formats are doable.

Potential Problems with Non-Traditional Item Types

The ICT measure under development presumably has the potential to be integrated into the current CAT-ASVAB framework. Characteristics of some of the non-traditional items presented in this report pose two potential problems for integration into the CAT-ASVAB system. Both issues are related to the measurement model currently in use. A fundamental assumption of Item Response Theory (IRT) is that, controlling for the latent construct being measured, there is no relationship between individual observed variables (i.e. test items). That is, once the effect of the latent trait being measured has been partialed out, the correlation between test items is zero. Hence, the probability of answering any one item correctly is independent from any other (Lord, 1980). This assumption is known as local independence and has implications for test item characteristics. It means that the construct of interest is wholly responsible for the relationship between test items, test items have no direct or indirect effect on one another, and that measurement errors associated with each item are uncorrelated (Bollen, 2002). Test items that are explicitly non-independent (i.e., pose multiple questions but refer to a common stimulus) will clearly violate this assumption. The CAT-ASVAB circumvents this issue by establishing functional independence between items in the Paragraph Comprehension subtest. Specifically, only one question is associated with each reading passage (Segall & Moreno, 1999). Although it would be possible to overcome this limitation in a similar manner, doing so greatly reduces the efficiency benefit of such items.

Polytomous scored item formats also present a challenge for integration with the CAT-ASVAB as it exists now. The CAT-ASVAB currently employs only dichotomously scored items using the 3 parameter logistic model (3PL; Lord & Novick, 1968). IRT models appropriate for polytomously scored items (e.g., Muraki, 1997) are available, and mixing of models is not problematic within the IRT framework per se. Nevertheless, the current CAT-ASVAB infrastructure is configured to work with the 3PL model only, and revising it to include other models would require substantial changes to the current system. The potential benefit of including polytomous items must be weighed against the costs associated with altering the current infrastructure.

References

Bliesener, T. (1996). Methodological moderators in validating biographical data in personnel selection. *Journal of Occupational and Organizational Psychology, 69*, 107–120.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of Psychology, 53,* 605-634.

Carlson, K. D., Scullen, S. E., Schmidt, F. L., Rothstein, H., & Erwin, F. (1999). Generalizable biographical data validity can be achieved without multi-organizational development and keying. *Personnel Psychology, 52,* 731–755.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Colberg, M. (1984). Towards a taxonomy of verbal tests based on logic. *Educational and Psychological Measurement*, *44*, 113-120.

Colberg, M. (1985). Logic based measurement of verbal reasoning: A key to increased validity and economy. *Personnel Psychology, 38,* 347-359.

Colberg, M., Nester, M. A., & Trattner, M. H. (1985). Convergence of the inductive and deductive models in the measurement of reasoning abilities. *Journal of Applied Psychology, 70,* 681-694.

Colberg, M. & Varon Cobos, M. C. (2000). *Taxonomy from the logical reasoning and quantitative reasoning subtests of the economists test.* Washington, D.C.: Bureau of Labor Statistics.

Dean, M. A. (2004). An assessment of biodata predictive ability across multiple performance criteria. *Applied H. R. M. Research,* 9, 1–12.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Drasgow, F., Olson-Buchanan, J. B., & Moberg, P. J. (1999). Development of an interactive video assessment: Trials and tribulations. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gandy, J. A., Dye, D. A., & MacLane, C. N. (1994). Federal Government selection: The individual achievement record. In G. S. Stokes, M. D. Mumford, and W. A. Owens (Eds.), *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction*. (pp. 275–309). Palo Alto, CA: CPP Books.

Gandy, J. A., Outerbridge, A. N., Sharf, J. C., & Dye, D. A. (1989). *Development and initial validation of the Individual Achievement Record (IAR)* (PRD-90-01). Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development.

Guilford, J. P., & Lacey, J. I. (Eds.) (1947). *Army Air Forces Aviation Psychology Program Research Reports: Printed classification tests* (Report No. 5). Washington, DC: U.S. Government Printing Office.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

Jensen, A.R. (1980). *Bias in mental testing.* New York: The Free Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F., M. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

Moss, F. A., & Hunt, T. (1926). Ability to get along with others. *Industrial Psychology*, 170–178.

Motowidlo, S., Dunnette, M., & Carter, G. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*, 640-647.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in Industrial Psychology*. Palo Alto, CA: Davies-Black Publishing.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J.A. Weekley & R.E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement and application.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Muraki E. (1997). A generalized partial credit model. In van der Linden, W. & Hambleton, R., *Handbook of modern item response theory.* New York: Springer.

Reilly, R. A., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–62.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). *Biographical data in employment selection: Can validities be made generalizable? Journal of Applied Psychology, 75*, 175–184.

Segall, D. O. & Moreno, K. E. (1999). Development of the CAT-ASVAB. In F. Drasgow & J. B. Olson-Buchanan (Eds.). *Innovations in Computerized Assessment* (pp. 35-65). Hillsdale, NJ: Lawrence Earlbaum Associates.

Simpson, R. W. & Nester, M. A. (2007). *Taxonomy for reasoning questions using logic based measurement*. Paper presented at the International Public Management Association for Human Resource Conference, St. Louis, MO.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item forms in computer-based testing: In persuit of improved construct representation. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development*, (pp. 329-347). Mahwah, NJ: Lawrence Earlbaum.

Stokes, G., Mumford, M., & Owens, W. (1994). *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction*. Palo Alto, CA, US: CPP Books.

Thorndike, R. L. (1936). Factor analysis of social and abstract intelligence. *The Journal of Educational Psychology*, XXVII, 231–233.

Waugh, G. W., & Russell, T. L. (2005). Predictor situational judgment test. In D. J. Knapp, C.E. Sager, & T.R. Tremble (Eds.) *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (TR 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, 52, 679–700.

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J.A. Weekley & R.E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement and application.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Wernimont, P., & Campbell, J. (1968). Signs, samples and criteria. *Journal of Applied Psychology*, *52*, 372-376.